Problem 2 (PCA, 15 pts)

For this problem you will implement PCA from scratch. Using `numpy` to call SVDs is fine, but don't use a third-party machine learning implementation like `scikit-learn`.

We return to the MNIST data set from T4. You have been given representations of 6000 MNIST images, each of which are 28×28 greyscale handwritten digits. Your job is to apply PCA on MNIST, and discuss what kind of structure is found.

The given code in `T5_P2.py` loads the images into your environment. File `T5_P2_Autograder.py` contains a test case to check your cumulative proportion of variance.

1. Compute the PCA. Plot the eigenvalues corresponding to the most significant 500 components in order from most significant to least. Make another plot that describes the cumulative proportion of variance explained by the first k most significant components for values of k from 1 through 500. How much variance is explained by the first 500 components? Describe how the cumulative proportion of variance explained changes with k.

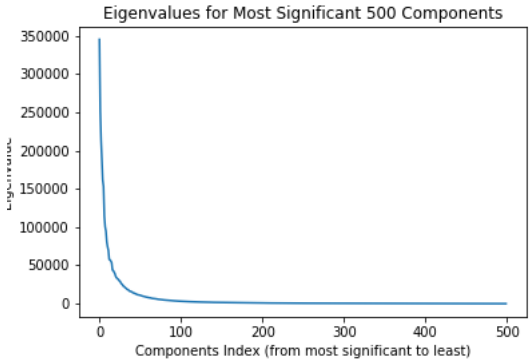
2. Plot the mean image of the dataset and plot an image corresponding to each of the first 10 principle components. How do the principle component images compare to the cluster centers from K-means? Discuss any similarities and differences. Include all 11 plots in your PDF submission.

3. Compute the reconstruction error on the data set using the mean image of the dataset. Then compute the reconstruction error using the first 10 principal components. How do these errors compare to the final objective loss achieved by using K-means on the dataset? Discuss any similarities and differences.

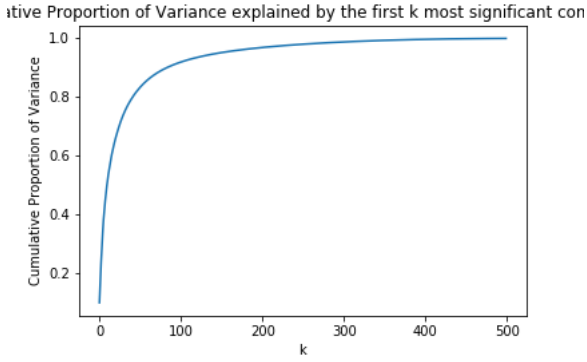
Include your plots in your PDF. There may be several plots for this problem, so feel free to take up multiple pages.

Solution

1. The plot of the eigenvalues corresponding to the most significant 500 components in order from most significant to least is shown below.

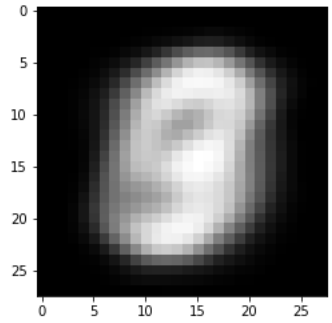


The plot of the cumulative proportion of variance explained by the first 500 most significant components is shown below.

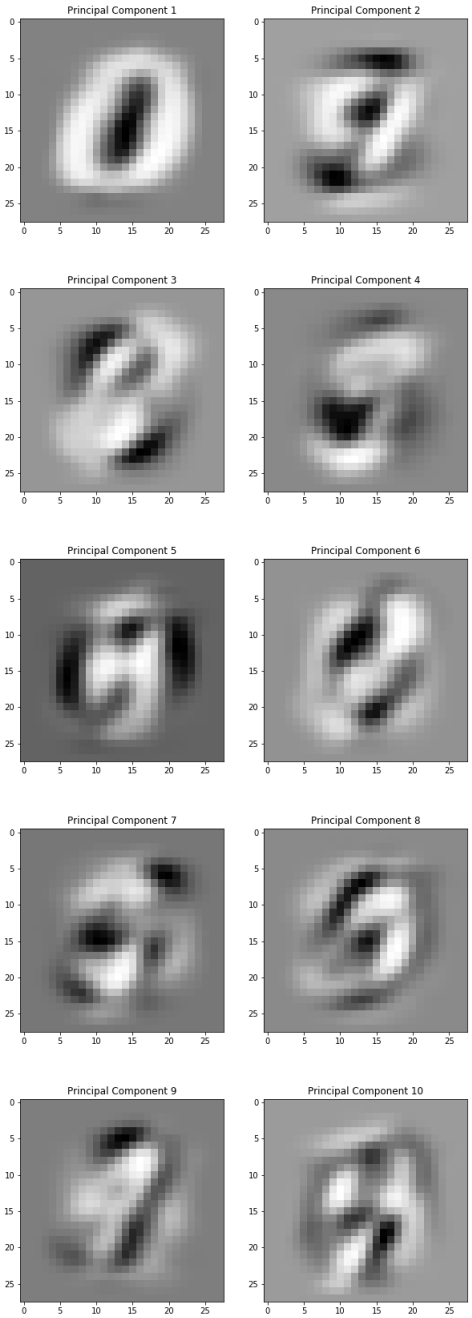


99.9398% of the total variance in the data is explained by the first 500 components. As k increases, the cumulative proportion of variance explained increases at a decreasing rate. At a value of k of about 80-100, the plot plateaus, showing that each additional component (increasing k by 1) does not significantly increase the amount of variance explained by the model. This is consistent with the plot of the eigenvalues, which shows that the eigenvalues past approximately the 80th most significant component are all approximately equal and much smaller than the eigenvalues for the first 80 most significant components, which implies that most of the variance in the model can be explained by just the first 80 most significant components.

2. The mean image of the dataset is plotted below.



The images for the first 10 principal components are shown below.


Principal Component 1
Principal Component 2
Principal Component 3
Principal Component 4
Principal Component 5
Principal Component 6
Principal Component 7
Principal Component 8
Principal Component 9
Principal Component 10

Comparing to the images of the 10 principal components to the cluster center images of the K Means from the Staff's Solution to Homework 4, we find that the 10 principal component images are most comparable to the standardized K Means images. In particular, the non-standardized K Means images are the easiest and most decipherable of the three sets of images, and the only set of images with a black background. As in homework 4, we find that the standardized K Means images appear harder to read and more indecipherable than the non-standardized K Means images. While most of the standardized K Means images are readable, some are so obscured that they are impossible to read. In contrast, while both the 10 principal component images and the standardized K Means images have gray backgrounds, the 10 principal component images are the most indecipherable of the three sets of images, and in fact only one or two of the 10 principal component images actually appear like readable digits. The intuition for this is that while the K Means centers are displaying mean images of each cluster, which will cause the image to look like a readable digit depending on how well the cluster is formed, the principal components do not actually depict digits, but rather the most significant features of digits such that if we combined some of the principal component images, the resulting image would appear to be a decipherable digit. This is the primary reason why the plots of the principal components do not appear to be like digits, since the principal components are not actually images, but rather the most significant features of the images.

3. The Reconstruction Error when using the mean image is: 3436023.412. The Reconstruction Error when using the first 10 principal components is: 1731603.929. For K Means, the Final Objective is: 2576769.813. The Reconstruction Error when using the mean image is larger than the Final Objective Loss when using K Means. However, the Reconstruction Error when using the first 10 principal components is smaller than the Final Objective Loss when using K Means. This makes intuitive sense. We expect that using the mean image yields the greatest error, since this is akin to using just one cluster in K Means or no components in principal components (since we recentered our data to have a mean of 0). Using 10 principal components has better performance than 10 clusters K Means since the 10 principal components gives us the 10 directions of largest variance in the data, whereas K Means just clusters the nearest points together. It is preferable to use the principal components because we are capturing the most significant sources of variation in our data and will therefore allow our model have a lower error than the K Means model, which simply seeks to just cluster similar images.